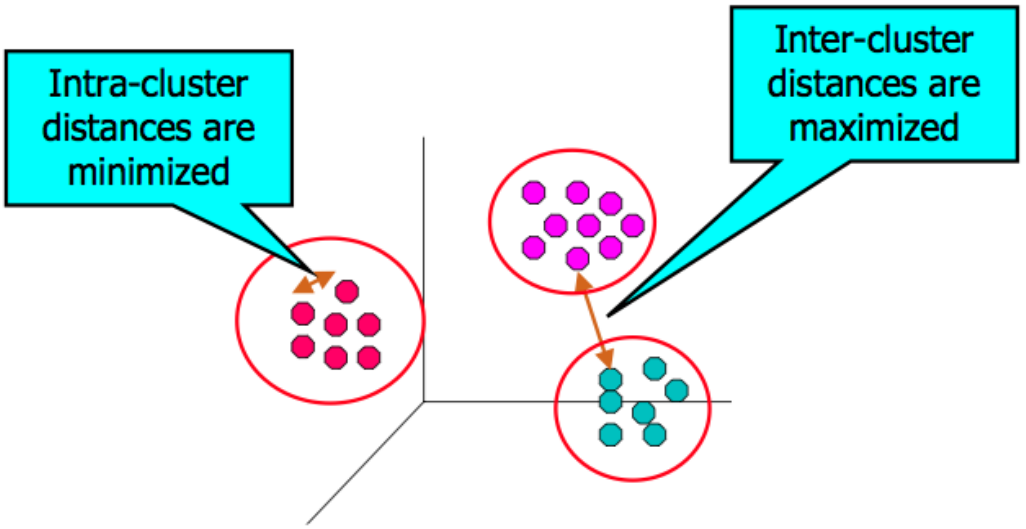
Analisis Cluster

Analisis Cluster

- Analisis cluster adalah pengorganisasian kumpulan pola ke dalam cluster (kelompok-kelompok) berdasar atas kesamaannya.
- Pola-pola dalam suatu cluster akan memiliki kesamaan ciri/sifat daripada pola-pola dalam cluster yang lainnya.



Analisis Cluster

- Clustering bermanfaat untuk melakukan analisis pola-pola yang ada, mengelompokkan, membuat keputusan dan machine learning, termasuk data mining, document retrieval, segmentasi citra, dan klasifikasi pola.
- Metodologi clustering lebih cocok digunakan untuk eksplorasi hubungan antar data untuk membuat suatu penilaian terhadap strukturnya.

Tipe Clustering

- Partitional Clustering
  - Pembagian objek data ke dalam non-overlapping subset (cluster) sehingga setiap objek data adalah tepat satu subset
- Hirerarchical Clustering
  - Sehimpunan cluster bersarang yang diorganisasikan sebagai struktur hirarki pohon.

Tipe Cluster

- Well-separated clusters
- Center-based clusters
- Density-based clusters

Well-separated

- Sebuah cluster adalah sehimpunan titik yang memiliki kemiripan dengan titik lain dalam cluster daripada di cluster lain.

Center-based

- Sebuah cluster yang memiliki anggota-anggota yang mirip dengan pusat cluster daripada pusat cluster lain.
- Pusat cluster
  - Centroid: Rata-rata dari semua titik dalam cluster
  - Medoid: memilih titik sebagi titik tengah.

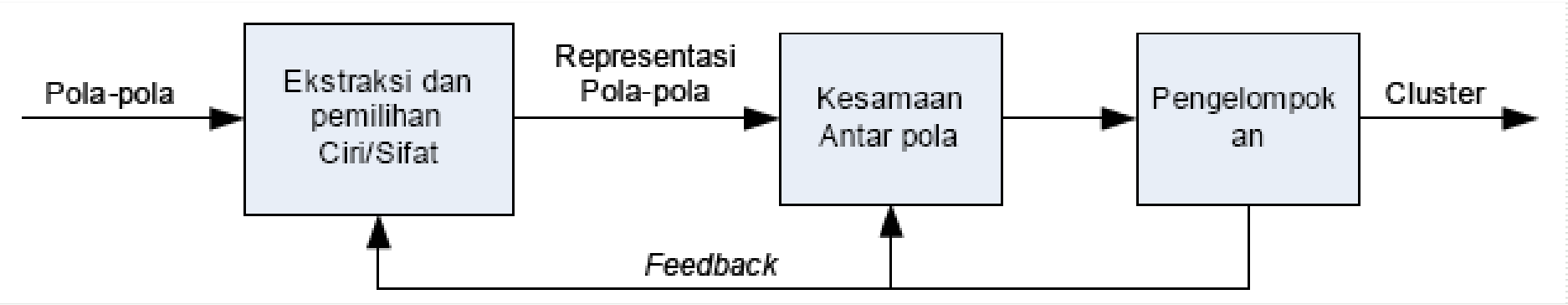
Density-based

- Sebuah cluster adalah area padat titik, yang dipisahkan dengan area kepadatan rendah, dari area kepadatan tinggi lainnya.
- Digunakan ketika cluster tidak teratur atau saling terkait, dan ketika noise dan outliers hadir.

Komponen

- representasi pola (termasuk ekstraksi sifat/ciri dan atau pemilihan),
- definisi ukuran kedekatan pola sesuai dengan domain data,
- clustering atau pengelompokan,
- jika diperlukan, abstraksi data (proses ekstraksi untuk deksripsi cluster),
- jika diperlukan, penilaian terhadap hasil (menggunakan metode pengukuran dan pengujian terhadap hasil clustering apakah valid atau tidak).

Tahapan Clustering



- Kedekatan pola biasanya diukur dengan fungsi jarak antar dua pasang pola.
  - *cosine similarity*, *manhattan distance*, dan *euclidean distance*.

Tahapan Clustering

- Representasi pola (pattern representation) merupakan jumlah kelas, jumlah pola yang ada, jumlah, tipe dan skala ciri/sifat yang tersedia untuk algoritma clustering.
- Pemilihan ciri/sifat (feature selection) adalah proses identifikasi ciri/sifat yang lebih efektif untuk digunakan dalam algoritma clustering, sedangkan ekstraksi ciri/sifat adalah pemakaian satu atau lebih transformasi dari ciri/sifat yang ada sebelumnya untuk mendapatkan ciri/sifat yang lebih menonjol.

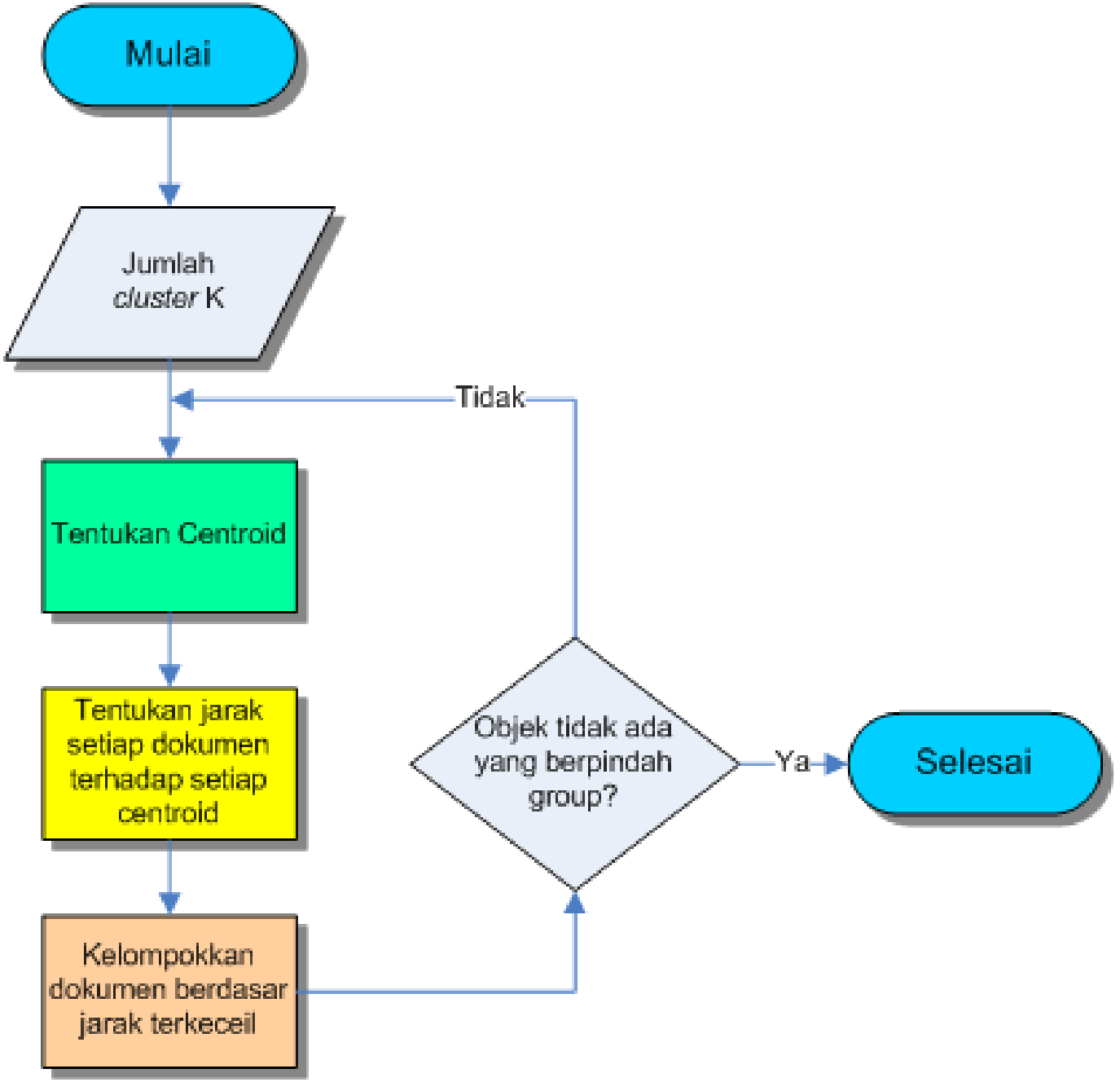
Tahapan Clustering

- Kedekatan pola biasanya diukur dengan fungsi jarak antar dua pasang pola.
- Pengukuran jarak yang sederhana, seperti *Euclidean distance*, *Minkowski*, *Hamming distance*, sering digunakan untuk menyatakan ketidaksamaan antara dua pola
- Sedangkan pengukuran kesamaan lain, seperti *Simple Matching Coefficient*, *Jaccard Coefficient*, *Cosine Similarity*, dapat digunakan untuk menunjukkan kesamaan karakter antar pola-pola.

k-Means

- Partitional clustering
- Setiap cluster terasosiasi dengan sebuah centroid
- Setiap titip dinyatakan ke suatu cluster yang paling dekat dengan centroidnya.
- Jumlah cluster, K, dinyatakan di awal

K-Means


Mulai
Jumlah cluster K
Tidak
Tentukan Centroid
Tentukan jarak setiap dokumen terhadap setiap centroid
Kelompokkan dokumen berdasar jarak terkeceil
Objek tidak ada yang berpindah group?
Ya
Selesai

Contoh K-Means

- Kelompokkan dataset berikut ke dalam 3 kelompok dengan k-means (2 epoch saja):
  - A1=(2,10)
  - A2=(2,5)
  - A3=(8,4)
  - A4=(5,8)
  - A5=(7,5)
  - A6=(6,4)
  - A7=(1,2)
  - A8=(4,9)

Keterbatasan K-Means

- K-Mean bermasalah ketika cluster-cluster berbeda
  - Ukuran
  - Kepadatan
  - Tidak berbentuk bola
- K-Mean bermasalah ketika data berisi outlier

K-Medoid

- Seperti metode partisi clustering yang lainnya, metode k-medoid juga digunakan untuk pengelompokkan dokumen.
- Dalam metode k-medoid ini setiap cluster dipresentasikan dari sebuah objek di dalam cluster yang disebut dengan medoid.
- Tujuannya adalah menemukan kelompok k-cluster (jumlah cluster) diantara semua objek data di dalam sebuah kelompok data.
- Clusternya dibangun dari hasil mencocokkan setiap objek data yang paling dekat dengan cluster yang dianggap sebagai medoid sementara.

K-Medoids

1. pilih point k sebagai inisial centroid / nilai tengah (medoids) sebanyak k cluster.
2. cari semua point yang paling dekat dengan medoid, dengan cara menghitung jarak vector antar dokumen. (menggunakan Euclidian distance)
3. secara random, pilih point yang bukan medoid.
4. hitung total distance
5. if TD baru < TD awal, tukar posisi medoid dengan medoids baru, jadilah medoid yang baru.
6. ulangi langkah 2 - 5 sampai medoid tidak berubah.

Contoh K-Medoids

X1	2	6
X2	3	4
X3	3	8
X4	4	7
X5	6	2
X6	6	4
X7	7	3
X8	7	4
X9	8	5
X10	7	6

K=2, c1(3,4); c2(7,4)

c1	Data objects (Xi)		Cost (distance)
3	4	2	6
3	4	3	8
3	4	4	7
3	4	6	2
3	4	6	4
3	4	7	3
3	4	8	5
3	4	7	6

c2	Data objects (Xi)		Cost (distance)
7	4	2	6
7	4	3	8
7	4	4	7
7	4	6	2
7	4	6	4
7	4	7	3
7	4	8	5
7	4	7	6

Nearest Neighbor clustering

- Sebuah titik membentuk cluster baru atau bergabung dengan salah satu cluster yang sudah ada bergantung pada seberapa dekat titik tersebut dengan cluster.
  - Sebuah treshold, t, untuk menentukan bergabung atau membuat cluster baru.

Nearest Neighbor clustering

Input:

$D=\{t_1, t_2, \ldots, t_n\}$ // Set of elements
A // Adjacency matrix showing distance between elements
θ // threshold

Output:

K //Set of k clusters

Nearest-Neighbor algorithm

$K_1 =\{t_1\}$; add K_1 to K; // t1 initialized the first cluster
k = 1;
for i =2 to n **do** // for t_2 to t_n add to existing cluster or place in new one
find the t_m in some cluster K_m in K such that $d(t_m,t_i)$ is the smallest;
if $d(t_m,t_i) < \theta$ **then**
$K_m = K_m \cup \{t_i\}$ // existing cluster
else
$k = k + 1$; $K_k = \{t_i\}$; add K_k to K // new cluster

Latihan NN

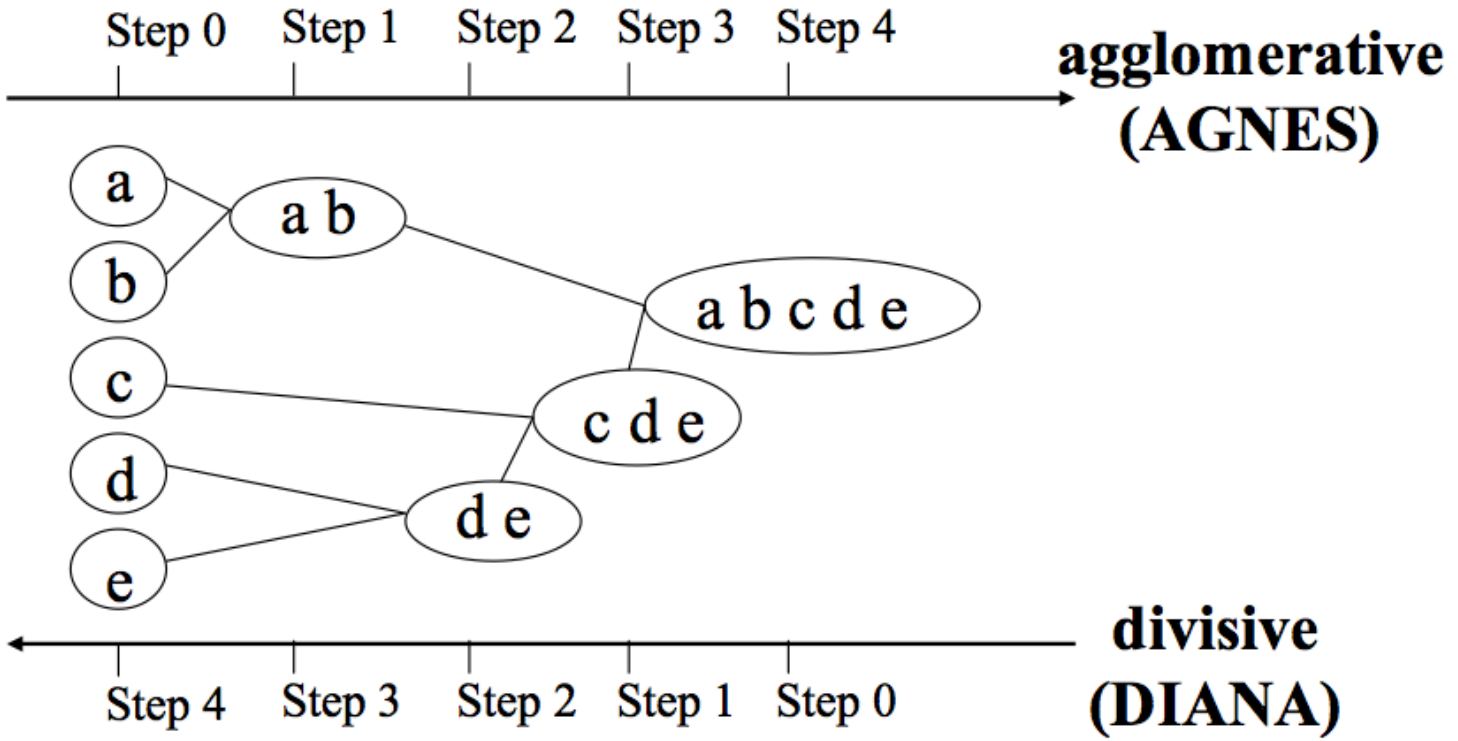
- Kelompokkan dataset berikut ke dalam 3 kelompok dengan NN clustering (2 epoch saja):
  - A1=(2,10)
  - A2=(2,5)
  - A3=(8,4)
  - A4=(5,8)
  - A5=(7,5)
  - A6=(6,4)
  - A7=(1,2)
  - A8=(4,9)

Hierarchical Clustering

- Membentuk beberapa himpunan cluster
  - Jumlah cluster tidak dimasukkan di awal
- Struktur hirarki cluster dapat dipresentasikan sebagai dendrogram.
  - Daun berisi 1 item.
  - Setiap item masuk dalam satu cluster
  - Root mewakili semua item
  - Internal node menyatakan cluster yang dibentuk oleh penggabungan cluster anak.
  - Setiap level diasosiasikan dengan suatu treshold jarak yang digunakan untuk menggabungkan cluster
    - Jika jarak antar 2 cluster lebih kecil dari treshold, maka digabungkan.
    - Jarak akan bertambah sesuai dengan level.

Hierarchical Clustering

- Menggunakan matrik jarak sebagai kriteria clustering. Metode ini tidak memerlukan jumlah cluster, K, sebagai inputan, namun butuh kondisi terminasi.



Single Link dan Complete Link

- Single Link
  - 2 cluster digabungkan jika hanya 2 titiknya berdekatan.
- Complete Link
  - Jarak antar 2 cluster adalah jarak terbesar antar sebuah elemen dalam satu cluster dan sebuah elemen di cluster lain.

Contoh: AGNES

	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

Single Link: AGNES

d	k	K
0	4	{A}, {B}, {C}, {D}
1	3	{A, B}, {C}, {D}
2	2	{A, B, C}, {D}
3	1	{A, B, C, D}

d	k	K
0	4	{A}, {B}, {C}, {D}
1	3	{A, B}, {C}, {D}
2	3	{A, B}, {C}, {D}
3	2	{A, B}, {C, D}
4	2	{A, B}, {C, D}
5	2	{A, B}, {C, D}
6	1	{A, B, C, D}

Complete Link: AGNES